Received: 3 May 2022 | Revised: 26 June 2022 | Accepted: 8 July 2022

DOI: 10.1111/cts.13371

ARTICLE

A novel analytical framework for risk stratification of real-world data using machine learning: A small cell lung cancer study

Luca Marzano[1] | **Adam S. Darwich**[1] | **Salomon Tendler**[2] | **Asaf Dan**[2] | **Rolf Lewensohn**[2] | **Luigi De Petris**[2] | **Jayanth Raghothama**[1] | **Sebastiaan Meijer**[1]

[1]Division of Health Informatics and Logistics, School of Engineering Sciences in Chemistry, Biotechnology and Health (CBH), KTH Royal Institute of Technology, Huddinge, Sweden

[2]Department of Oncology-Pathology, Karolinska Institutet and the Thoracic Oncology Center, Karolinska University Hospital, Stockholm, Sweden

Correspondence
Luca Marzano, Division of Health Informatics and Logistics, School of Engineering Sciences in Chemistry, Biotechnology and Health (CBH), KTH Royal Institute of Technology, Huddinge, Sweden.
Email: lmarzano@kth.se

Funding information
Stockholm Cancer Society, Grant/Award Number: #201103, #174063 and #201202; Swedish Cancer Society, Grant/Award Number: CAN2021/1469 Pj01 and CAN 2018/597; KTH Royal Institute of Technology

Abstract

In recent studies, small cell lung cancer (SCLC) treatment guidelines based on Veterans' Administration Lung Study Group limited/extensive disease staging and resulted in broad and inseparable prognostic subgroups. Evidence suggests that the eight versions of tumor, node, and metastasis (TNM) staging can play an important role to address this issue. The aim of the present study was to improve the detection of prognostic subgroups from a real-word data (RWD) cohort of patients and analyze their patterns using a development pipeline with thoracic oncologists and machine learning methods. The method detected subgroups of patients informing unsupervised learning (partition around medoids) including the impact of covariates on prognosis (Cox regression and random survival forest). An analysis was carried out using patients with SCLC ($n = 636$) with stage IIIA–IVB according to TNM classification. The analysis yielded $k = 7$ compacted and well-separated clusters of patients. Performance status (Eastern Cooperative Oncology Group-Performance Status), lactate dehydrogenase, spreading of metastasis, cancer stage, and CRP were the baselines that characterized the subgroups. The selected clustering method outperformed standard clustering techniques, which were not capable of detecting meaningful subgroups. From the analysis of cluster treatment decisions, we showed the potential of future RWD applications to understand disease, develop individualized therapies, and improve healthcare decision making.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

Real-world data (RWD) has the potential to inform real-world evidence (RWE) and significantly impact clinical practice as well as clinical trial design.

WHAT QUESTION DID THIS STUDY ADDRESS?

Can practice-based patient data be used to stratify patient groups to predict clinical treatment pathways and prognosis? How can RWD challenges be addressed to achieve RWE of relevance to treatment decisions?

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

There are a series of aspects to take into account when RWD are adopted in clinical studies, to avoid biases and produce reliable results. Clinical experts play a key role regarding clinical inference and unfolding of treatment process insights contained within these datasets.

HOW MIGHT THIS CHANGE DRUG DISCOVERY, DEVELOPMENT, AND/OR THERAPEUTICS?

Meaningful risk stratification from RWD could improve knowledge of disease, design individualized treatment pathways, and advise the study design of future clinical trials. The study provides insights into the unique challenges of RWD and how these can be addressed.

INTRODUCTION

Stratification of patients to predict accurate prognostic outcomes and improve treatment selection is a crucial challenge in oncology.[1] Increasing availability of real-world data (RWD) provides the opportunity to aid grouping based on prognostic indicators[2] and fill the knowledge gap between randomized controlled trials (RCTs) and everyday clinical practice, as well as inform study design (e.g., emulation of control arms).[3–5]

Small cell lung cancer (SCLC) is an ideal case for performing risk stratification from clinical RWD. Prediction of clinical outcomes is challenging due to the rapid formation of multiple distant metastases and the lack of knowledge regarding chemo-resistance mechanisms.[6,7] In fact, advancements in treatment strategies have been limited over the past 30 years.[8]

Treatment selection has historically been based on the Veterans' Administration Lung Study Group (VALSG) staging, which includes limited disease (LD), the tumor is confined to one hemithorax, which is usually treated with a combination of chemotherapy and thoracic radiotherapy (1.5 Gy fraction up to a total dose of 45 Gy) and extensive disease (ED), the tumor has spread in other parts of the body, where treatment is limited to palliative chemotherapy.[9] LD/ED staging is still widely used, nevertheless, this categorization is broad and unable to sufficiently differentiate prognostic subgroups.[10] This, even though considering that a substantial number of patients are in a far too poor general condition to tolerate any oncologic therapy at the time of diagnosis. In recent studies, it has been shown that the eighth version of the International Association for the Study of Lung Cancer (IALSC) tumor, nodes, and metastasis (TNM) staging was superior to VALSG staging for the assessment of patients' prognosis.[10,11]

However, RWD poses a series of practical challenges, including data quality, sample size, defining role of variables in clinical processes, addressing missing values and potential biases, and interpretation of results.[2,12–14] Furthermore, censoring of data should be handled in risk stratification tasks.[15]

Cox hazard ratios are widely used in oncology for the comparison of survival outcomes.[15,16] However, these are not capable of providing global comparisons between survival outcomes.[16] Moreover, these models rely on the assumption that any covariate effects on hazard are linear. This makes it difficult to make clinical inferences from retrospective studies. Machine learning applied to survival analysis has recently been proposed to overcome these limitations.[17,18]

The aim of this work was to explore clinical treatment patterns and outcomes relative to baseline patient characteristics in patients with SCLC. Here, we proposed a novel approach that merges statistical analysis with machine learning techniques for survival analysis (Random Survival Forest [RSF][19]), and informs unsupervised learning (Gower similarity[20] with Partition Around Medoids [PAM][21]) with the prognostic impact of covariates, thus detecting clinical meaningful subgroups.

Thus far, these methods have not been applied to healthcare RWD of patients with SCLC. To the authors knowledge, the present study is the first in which the combination of survival analysis and unsupervised machine learning resulted in a comprehensive separation of SCLC prognostic groups. Furthermore, the pipeline was designed to address several RWD challenges by including clinical experts to understand the treatment process and interpret the results.

MATERIALS AND METHODS

Cohort description

Part of the present cohort was previously used to validate the eighth TNM system and explore clinical outcomes in SCLC.[10,22] The cohort consisted of consecutive cases diagnosed and treated at Karolinska University Hospital between 2008 and 2016. All patients had previously been

reclassified from the older VALSG classification system to the eighth edition of the TNM. The study was approved by the institutional review boards at Karolinska Institutet and at Stockholm County Council (2016/8-31). The baselines included in the study were: TNM tumor (*T8*), nodes (*N8*), and metastasis (*M8*) cancer stage descriptors proposed by the IASLC with the relative stage (*ST8*), age, gender, Eastern Cooperative Oncology Group Performance Status (ECOG-PS), smoking status, positron emission tomography/computed tomography scan and brain computed tomography scan. Furthermore, hematology and blood chemistry values before the initiation of first-line treatment were obtained, this included CRP (mg/L), lactate dehydrogenase (LDH; μkat/L), albumin (g/L), sodium (Na; mmol/L), and hemoglobin (HB; g/L). LDH and CRP distributions were highly skewed, therefore, a log transformation was applied to these variables.

Missing laboratory values in the retrospective dataset (35% of the patients were missing LDH, CRP, HB, Na, and albumin) were imputed using the missing random forest algorithm.[23] This method showed better performances compared to other missing imputation techniques in previous studies (Supplementary Section S2). Patients with stage IIIA–IVB ($n = 636$) were included in the analysis (see Table S1). Patients that received radiotherapy or surgery alone, as well as patients with ECOG-PS = 4 were excluded ($n = 30$).

The conceptual treatment decision process was reconstructed based on the data and clinical expert feedback (Figure 1). Contraindications for receiving treatment included advanced age, significant comorbidities, and ECOG-PS 3–4. Moreover, some patients chose to decline or discontinue treatment due to the anticipated risk of side effects. Irradiation could be contraindicated for patients with LD with a large tumor burden or ECOG-PS 3–4.

The recommended treatment combination for SCLC is four cycles of platinum doublet chemotherapy (cisplatin or carboplatin together with etoposide or irinotecan). Platinum re-challenge with etoposide or irinotecan is recommended for treating relapse of a tumor that initially was chemotherapy sensitive but shows disease progression (after 180 days from start or 90 days after the end of the treatment), whereas re-challenge with platinum and irinotecan, or non-platinum regimens (monotherapy with etoposide, irinotecan, or topotecan) are administrated during the second line for refractory relapse (progression of the tumor burden before 180 days from start or 90 days from the end of initial treatment).[24]

Patients achieving a major tumor shrinkage after primary therapy, and who are still in good general condition (according to ECOG-PS) might be offered prophylactic cranial irradiation (PCI) to reduce the risk of developing brain metastases.[25,26]

The following variables related to treatment decisions and survival outcomes were included (Table 1): death (Censor = No) or censoring (no information available at the end of the study, Censor = Yes), overall survival (OS; days from the start of the oncologic therapy until the

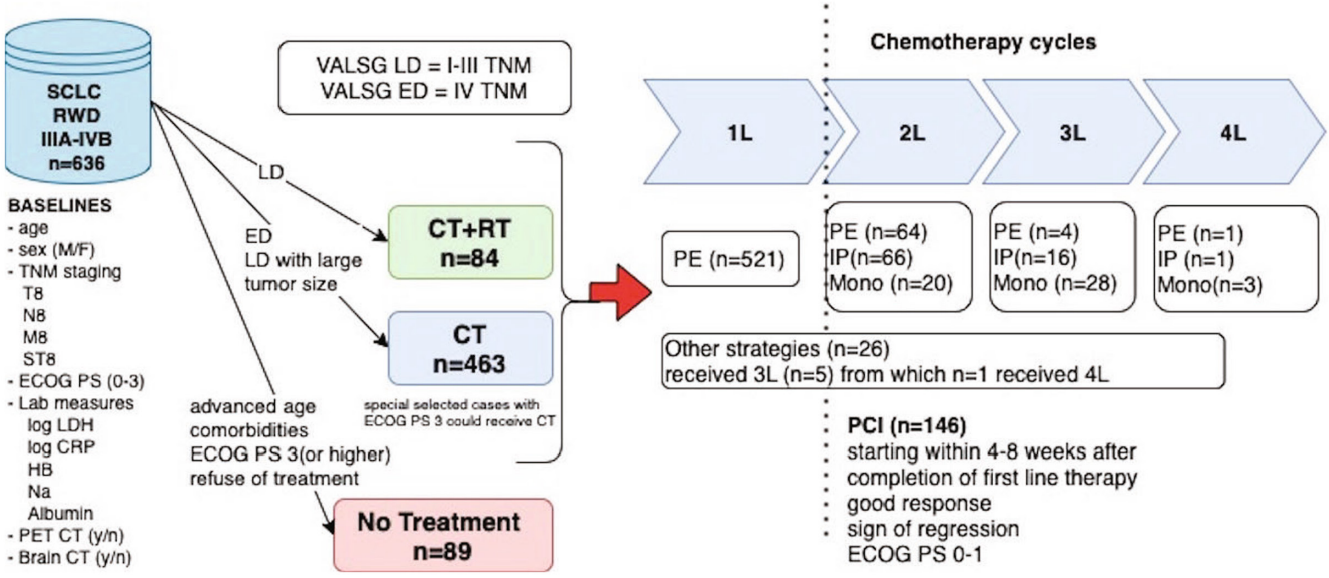

FIGURE 1 Treatment decision process for the cohort. CT, chemotherapy; CT + RT, chemotherapy and radiotherapy; ECOG, Eastern Cooperative Oncology Group; IP, platinum with irinotecan, monotherapy: etoposide, irinotecan or topotecan without platinum; LDH, lactate dehydrogenase; PCI, prophylactic cranial irradiation, PE, platinum with etoposide; PET, positron emission tomography; RT, radiotherapy; RWD, real-world data; SCLC, small cell lung cancer; TNM, tumor, node, metastasis. Others: treatment strategies where PE was not administrated at the first-line treatment.

TABLE 1 Treatment decision and follow-up outcomes

	No treatment ($N = 89$)	CT + RT No PCI ($N = 12$)	CT + RT PCI ($N = 72$)	CT No PCI ($N = 389$)	CT PCI ($N = 74$)
OS (days)					
Mean (SD)	37.9 (46.3)	592 (667)	1280 (829)	224 (194)	830 (587)
Median [min, max]	24.0 [1.00, 266]	363 [69.0, 2470]	1080 [232, 3500]	201 [4.00, 1360]	618 [321, 3730]
Censor					
No	89 (100%)	10 (83.3%)	44 (61.1%)	389 (100%)	68 (91.9%)
Yes	0 (0%)	2 (16.7%)	28 (38.9%)	0 (0%)	6 (8.1%)
Treatment					
No treatment	89 (100%)				
PE		6 (50.0%)	41 (56.9%)	298 (76.6%)	26 (35.1%)
PE-IP		3 (25.0%)	7 (9.7%)	29 (7.5%)	27 (36.5%)
PE-monotherapy		0 (0%)	1 (1.4%)	15 (3.9%)	4 (5.4%)
PE-PE		2 (16.7%)	20 (27.8%)	27 (6.9%)	15 (20.3%)
Others		1 (8.3%)	3 (4.2%)	20 (5.1%)	2 (2.7%)

Abbreviations: Censor, death (no) or censorship (yes) after OS days; CT, chemotherapy; CT + RT, chemotherapy and radiotherapy; OS, overall survival; PCI, prophylactic cranial irradiation; PE, no other lines were administrated after the first cycles PE; PE-PE, rechallenge with the same therapy; PE-IP, re-challenge with platinum and irinotecan; PE-monotherapy, re-challenge with monotherapy. Other treatment strategies where PE was not administrated at first-line were labeled as others.

occurrence of the event censoring/death), first-line therapy and PCI yes/no. In the studied cohort, only 53 patients received the third-line therapy, and very few subjects received a fourth-line treatment ($n = 6$). For this reason, only the administrated agents of the first two chemotherapy cycles were considered in Table 1.

Prognostic clustering pipeline

Figure 2 details the modeling workflow. The cohort was clustered into groups of mutually exclusive patients using the clinical characteristics in Figure 1, then clinical and treatment patterns were analyzed (Table 1). The method used in this study was as follows:

1. Covariate selection: a multivariate Cox proportional hazards regression was performed with clinical baselines described in Figure 1. Covariates having statistically significant hazard ratios according to Wald's test ($p < 0.05$) were selected.
2. Covariate importance: RSF was performed on the selected covariates. Feature importance was computed for all baselines with the permutation method.
3. Prognostic distance: weighted Gower similarity was computed with the selected covariates. The weight of each covariate in the distance computation was the feature importance obtained in the previous step.
4. Unsupervised machine learning: subgroups of patients were detected with the defined prognostic distance as input of the Partition Around Medoids (PAM) algorithm analysis of detected subgroups: detected groups of patients by PAM were analyzed with thoracic oncologists. Patterns of selected clinical baselines and treatment decisions were explored (Table 1).

Theoretical description of models in the pipeline is reported in Supplementary Sections S4–S6.

Clinical impact of SCLC covariates on prognosis have previously been studied with Cox regression.[8,27–29] In fact, the previous work with this cohort explored hazard ratios of patients with LD/ED receiving chemotherapy.[22] RSF was adopted in this work to compute feature importance of covariates that provided statistical significant hazard ratios. RSF is an extension of Random Forest suited for time-to-event analysis.[19] RSF was chosen for its capability of generalization and consistency of performance,[30] and its previous applications on SCLC radiological features and genetic data.[31,32]

Feature importance was computed according to the method presented by Fisher et al.[33] This was done to increase the interpretability of the machine learning model, and allow quantification of global information on the impact of covariates on prognosis. Given a covariate, its feature importance was computed by generating random permutations, computing the performance of the model.

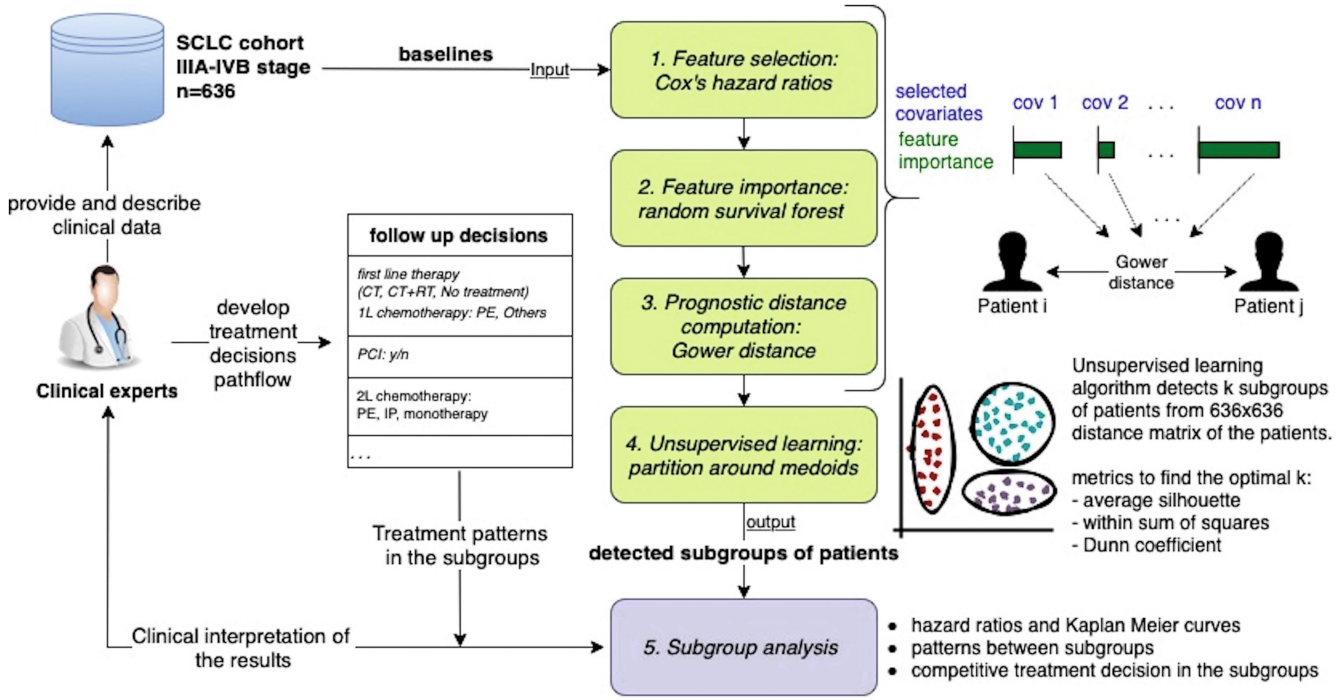

FIGURE 2 Data analysis and model development framework. The aim was to compute a pairwise similarity measure with the most relevant clinical baselines according to their prognosis (steps 1, 2, and 3), group patients with similar characteristics (step 4), and finally, study treatment patterns (Figure 1) of detected groups with thoracic oncologists (step 5). CT, chemotherapy; IP, platinum with irinotecan; PE, platinum with etoposide; RT, radiotherapy; SCLC, small cell lung cancer.

Therefore, covariates with higher feature importance correspond to a higher impact on predicted prognosis. The sequential application of Cox and RSF allowed the selection of covariates with prognostic significance based on their levels (hazard ratios), and compare the relative impact between them (feature importance).

The pairwise distance between patients was computed with Gower distance to handle the different formats of potential covariates (continuous, binary, and categorical variables). Gower distance was previously explored to analyze treatment selection in an RWD cohort with patients with non-small cell lung cancer.[34] The RSF feature importance was assigned as covariate contribution in the distance processed by the unsupervised algorithm.

PAM is a partitioning clustering method similar to the k-means algorithm that also works with non-Euclidean distances, such as the Gower distance.

Instead of computing centroids, the algorithm assigns a cluster center based on one of the observations (the patient for whom the sum of all distances to the other patients in the cluster is minimal). These central patients of the cluster are called medoids.[21] Applications of unsupervised learning to SCLC have mostly adopted hierarchical clustering.[35–37] However, in recent studies with generic cohorts of patients with lung cancer, partition-based algorithms showed better performance as compared to hierarchical clustering.[38] The same cohort was grouped using PAM and hierarchical clustering to consider the seventh version of TNM staging, age, and histology of the cancer.[39] Clustering of patient subgroups was performed using PAM because of its robustness to noise and outliers.[40]

The optimal number of clusters k was chosen according to within sum of squares, average silhouette, and Dunn coefficient.[40]

We compared our method with traditional PAM and hierarchical clustering (no feature importance step), and hierarchical clustering with feature importance weights.

Univariate survival analysis was performed to assess the prognostic difference within the detected groups (hazard ratios and Kaplan Meier curves). Clinical characteristics were explored along with treatment patterns and associated medoids. Kruskal-Wallis test was performed to estimate statistical difference of OS.

Treatment decision bias was tested by running the analysis a second time, including PCI and first-line therapy in step 1 with the baselines in Figure 1.

RESULTS

The optimal configuration found was with $k = 7$ clusters (Figures 3 and S1). In Tables 2 and S2, the common

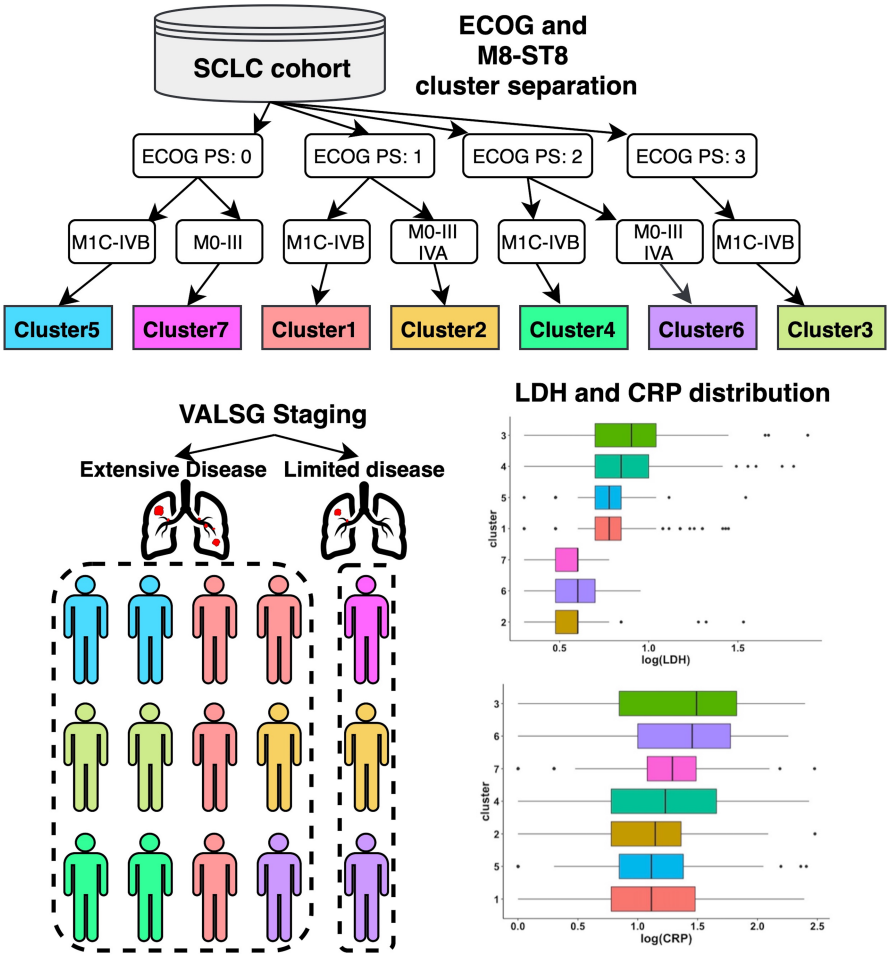

FIGURE 3 Prognostic clustering results. (a) Cox hazard ratios of selected covariates. (b) Feature importance distribution on 100 iterations of RSF (100 trees). (c) ECOG-PS-M8-ST8 cluster separation, LDH, and CRP distribution, and comparison with VALSG staging. (d) Univariate cluster's Cox hazard ratio. ECOG-PS, Eastern Cooperative Oncology Group-Performance Status; LDH, lactate dehydrogenase; RSF, Random Survival Forest; SCLC, small cell lung cancer; VALSG, Veterans' Administration Lung Study Group.

clinical features of the clusters are summarized along with treatment decisions and associated medoid characteristics. According to the Kruskal-Wallis test, OS distribution among clusters was statistically significant ($p < 2.2e^{-16}$).

According to feature importance, LDH and ECOG-PS were the most relevant for the prediction of OS (Figure 3b).

ECOG-PS was important in defining the patient subgroups. Figure 3c shows how the clusters were separated according to ECOG-PS M8-ST8 variables, LDH distribution, and the comparison with VALSG staging. Interestingly, 49 patients among the total 78 with stage IVA were associated to the clusters with patients having stage III rather than IVB.

Figure 4 shows the timeline of first-line treatment to illustrate competitive decisions in the clusters. CT and CT+ radiotherapy (RT) were competitive treatment decisions for cluster2 and cluster7 as well as CT and no treatment for cluster3 and cluster4. CT+ RT was the predominant treatment received by patients in cluster7 ($n = 50$), whereas most of the patients that did not receive treatment were in cluster3 ($n = 43$). The remaining clusters mainly included patients receiving CT.

Survival profiles were examined using univariate Cox regression (Figure 3d) and Kaplan Meier curves (Figure S3). Cluster7 and cluster2 showed the best prognosis, whereas cluster4 and cluster3 had the worst prognosis. Cluster1, cluster5 (patients treated with CT having ECOG-PS 0–1 and stage IVB) and cluster6 (patients treated with CT, having ECOG-PS 2 and stage III-IVA) had similar prognosis.

Treating patients with CT alone resulted in worse prognosis in cluster2 and cluster7 compared to CT+ RT. As expected, treating with CT resulted in better prognosis compared to patients that did not receive treatment in cluster3 and cluster4. Age between different treatment decision arms inside these clusters did not statistically differ.

A majority of patients received only one line of platinum with etoposide (PE). Subgroups of patients that received re-challenge with platinum (PE-PE and PE-platinum with irinotecan) were found in cluster1, cluster2, cluster5, and cluster7 (Tables S3, S4). However, inference about second-line treatment was not possible because of small sample sizes. In fact, there were no relevant patterns for PE-monotherapy, and the few patients receiving third line (3L) or fourth line (4L) treatment.

Patients that received PCI showed a better prognosis within their clusters. Comparison of survival profiles between the clusters led to similar results of previous LD/ED comparative studies (better OS for patients receiving PCI in cluster7 and cluster2 compared to cluster1 and cluster5).[41]

TABLE 2 Cluster patterns and medoids

Cluster	Clinical features	Treatment patterns	Medoid
Cluster1 ($n = 140$)	ECOG-PS: 1 TNM: M1C-IVB (132) LDH = 0.778 [0.699, 0.845] CRP = 1.114 [0.778, 1.48]	Treatment: CT (124) PCI (21) 2 L: PE-IP (19), PE-PE (12) 3 L (13)	TNM: T4N3M1C-IVB treatment: CT PE OS = 156 days
Cluster2 ($n = 91$)	ECOG-PS:1 TNM: M0-III (68) from which IIIA (18), IIIB (22) and IIIC (28), and IVA (23) LDH = 0.602 [0.477, 0.602] CRP = 1.146 [0.778, 1.362]	Treatment: CT (61) and CT + RT (27) PCI (47) 2 L: PE-PE (15), PE-IP (12) 3 L (13) and 4 L (2)	TNM: T4N3M0-IIIC treatment: CT + RT PCI PE-PE OS = 723 days
Cluster3 ($n = 97$)	ECOG-PS: 3 TNM: M1C-IVB (78) LDH = 0.903 [0.699, 1.041] CRP = 1.491 [0.845, 1.826]	Treatment: CT (54) and No treatment (43)	TNM: T4N2M1C-IVB treatment: CT PE OS = 17 days
Cluster4 ($n = 87$)	ECOG-PS: 2 TNM: M1C-IVB (84) LDH = 0.845 [0.699, 1] CRP = 1.23 [0.778, 1.658]	Treatment: CT (69) and No treatment (18)	TNM: T4N2M1C-IVB treatment: CT PE OS = 190 days
Cluster5 ($n = 85$)	ECOG-PS: 0 TNM: M1C-IVB (78) LDH = 0.778 [0.699, 0.845] CRP =1.114 [0.845, 1.38]	Treatment: CT ($n = 82$) PCI (11) 2 L: PE-IP (18), PE-PE (9) 3 L (10) and 4 L (1)	TNM: T4N2M1C-IVB treatment: CT PE-PE OS = 302 days
Cluster6 ($n = 46$)	ECOG-PS: 2 (39) and 3 (7) TNM: M0-III (29) from which IIIA (8), IIIB (11) and IIIC (10), and IVA (17) LDH = 0.602[0.477, 0.699] CRP = 1.455[1, 1.774]	Treatment: CT (34)	TNM: T4N3M0-IIIC treatment: CT PE OS = 503 days
Cluster7 ($n = 90$)	ECOG-PS: 0 TNM: M0-III (80) from which IIIA (27), IIIB (30), and IIIC (23) LDH = 0.602 [0.477, 0.602] CRP = 1.29 [1.079, 1.488]	Treatment: CT + RT (51) and CT (39) PCI (57) 2 L: PE-PE (20) and PE-IP (11) 3 L (14) and 4 L (3)	TNM: T4N2M0-IIIB treatment: CT PE OS = 1033 days

Note: For continuous variables, median [interquartile range] is reported. (2 L): second treatment line agent received. (3 L) and (4 L): number of patients that received further treatment lines after second.

Abbreviations: CT, chemotherapy; ECOG-PS, Eastern Cooperative Oncology Group-Performance Status; LDH, lactate dehydrogenase; OS, overall survival; PCI, prophylactic cranial irradiation; PE-IP, platinum with irinotecan; RT, radiotherapy; TNM, tumor, node, metastasis.

The approach outperformed standard hierarchical clustering, and PAM without feature importance weights in terms of clustering performance and novel detected groups (Figure S2).

DISCUSSION

In this paper, we showed the potential of using risk stratification based on RWD healthcare data collected over an extended time period. To the best of our knowledge, this is the first study of SCLC healthcare data using a survival machine learning model to detect subgroups with unsupervised learning informed by the prognostic impact of covariates.

The devised model development pipeline provided methodological solutions to address some of the highlighted challenges of using RWD.[2,12–14] The solution proposed to address RWD challenges and aspects that required further investigations are summarized in Table 3.

Figure 1 highlights the importance of contextualizing the clinical processes behind the data.[14] This allowed to avoid a strategy to reduce the risk stratification that was biased by treatment decisions (such as including PCI as an input covariate).

Continued discussions and iterative modeling with thoracic oncologists were instrumental to ensure that patients were clustered based on relevant baseline information, and to avoid potential biases or redundancy of covariates and treatment decisions.[1,42]

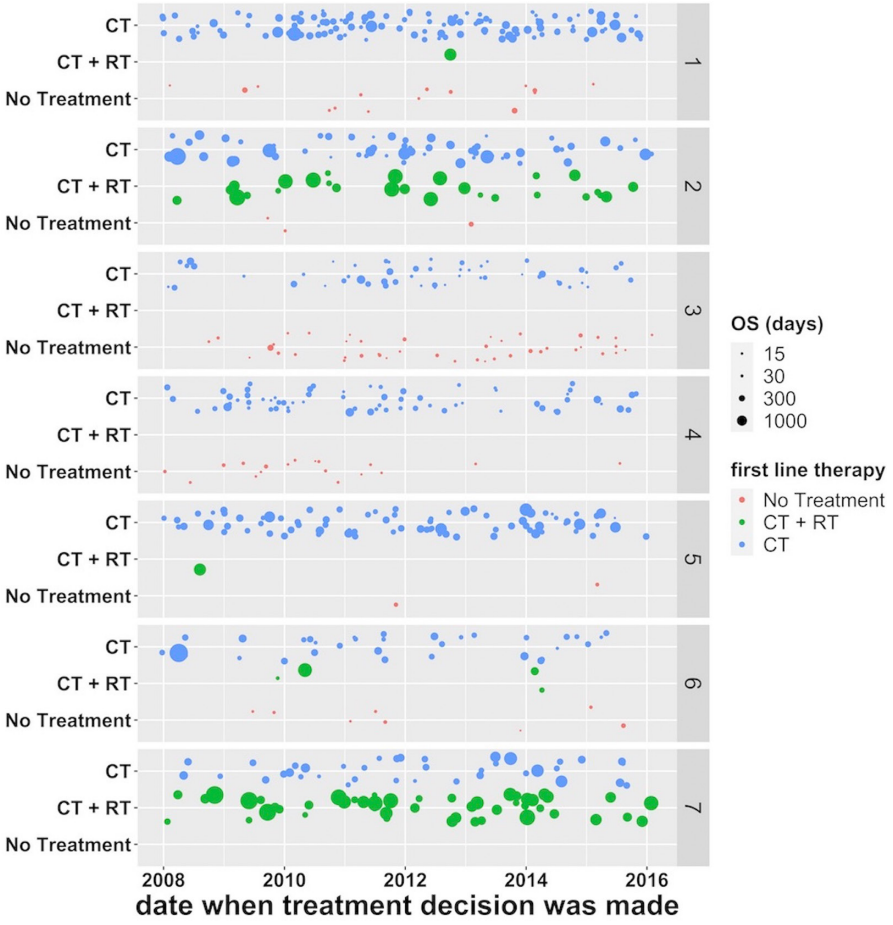

FIGURE 4 Treatment decisions in the cohort stratified by detected clusters across the years of diagnosis (2008–2016). CT, chemotherapy; RT, radiotherapy.

Previous studies grouped, a priori, the patients with LD-CT+RT and ED-chemotherapy according to VALSG.[22,28,43] In contrast, the approach presented here stratified the cohort a posteriori providing insights on staging patterns and competitive treatment decisions, thus extending the traditional pipeline of retrospective studies beyond the Cox Proportional Hazard model.

The covariates that characterize the clusters in Table 2 were consistent with findings of previous studies.[10,27]

This cohort is the most comprehensive re-staging of the eighth TNM version performed on patients with nonsurgical SCLC.[10] Moreover, the conclusions from the previous study on this cohort showed that the patient characteristics corresponded with historical data.[22] Hence, this cohort provided the opportunity for a robust external validation of earlier work.

Feature importance[33] allowed the estimation of global covariate effects, avoiding correlation bias or sample size effects on hazard ratios (e.g., the irrelevant effect of brain CT in Figure 3b).

The results support the importance of the prognostic impact of LDH and ECOG-PS.[28,29,44,45] Interestingly, LDH provided a higher feature importance score.

Using a combined survival analysis and unsupervised machine learning approach led to a comprehensive separation of SCLC prognostic groups (Figure 3c and Table 2). The approach outperformed traditional unsupervised techniques. The cluster analysis pointed out interesting considerations regarding the new stage categories IIIC, IVA, and IVB of the eighth version of TNM and how these are grouped.[46]

Figure 4 highlights the possibility to use longitudinal RWD to extract treatment arms of specific patient clusters to emulate clinical trials.[5] We found interesting competitive processes inside the clusters (CT+RT/CT and CT/no treatment). There were no relevant differences in treatment decisions over the considered years within the clusters. This allowed us to exclude the possibility of confounded analysis by differences in treatment modalities thought at the time.

Another interesting result was the similar survival profiles of CT therapy between patients with stage IVB and good ECOG-PS, and patients with stage III–IVA and worse ECOG-PS (cluster1, cluster5, and cluster6).

The analysis highlights the role of RWD to inform and shape future clinical trials.[2,5] Several of the findings were supported by previous RCTs in similar patient cohorts (e.g., higher OS for LD receiving CT+RT vs. CT alone).[41,47] On the other hand, the current work emphasized new patient subgroups and treatment outcomes in the different clusters (e.g., CT/no treatment in cluster3 and cluster4,

TABLE 3 Proposed solutions to address RWD challenges and future research to overcome remaining challenges

RWD challenges	Examples from the SCLC case	Implemented solutions	Identified solutions for future studies
Data quality	Retrospective clinical data Re-staging performed during 2008–2016 – lack of longitudinal information regarding treatment decision	Data pre-processing Multiple missing imputation	Collection of longitudinal information Include more recent patients' data
Sample size effects	Most of patients have advanced stage and multiple distant metastasis (M1C-IVB)	Synergy of survival analysis and unsupervised learning Separated patients in a meaningful way (otherwise main stage for all clusters would have been IVB)	Include more patients and increase sample size of limited subgroups (e.g., IVA stage)
Reconstruction of processes behind the acquired data	Censored data Correlation between covariates and treatment decision	Clinical experts in the loop: conceptualization of the treatment process Survival analysis to handle censoring	Enrichment of the process description through longitudinal variables Multistate models
Missing values	Laboratory values difficult to retrieve or not collected because of patient condition	Missing imputation with missForest algorithm	Lack of ground truth regarding the imputed values. Further research required
Biases and confounders	Baselines and treatment decision covariates Nonlinear effects Lack of longitudinal information	Separation of baselines, a posteriori study of treatment decisions Explainable machine learning (feature importance)	Study of time-dependent effects Study with cohorts from other centers
Interpretation of results	Cautious clinical inference regarding the results	Clinical interpretation of survival and cluster analysis Explainable machine learning (feature importance) Comparison with traditional unsupervised learning	Study of time-dependent effects Local explainable machine learning Sensitivity analysis comparing different strategies (covariate selection, survival and unsupervised algorithms, other cohorts and case studies)

Abbreviations: RWD, real-world data; SCLC, small cell lung cancer.

and the similar OS for patients in cluster1, cluster5, and cluster6).

This work also highlighted some limitations in the approach of generating RWE from RWD. Clinical inference should be made carefully due to the lack of further information regarding the patients, longitudinal measures, and treatment decisions (e.g., comorbidities, dosing, reason for cessation of treatment, and side effects).

Another limitation of the data is the unbalanced presence of patients with stage IVB disease that makes it challenging to apply the model to other patient cohorts with limited sample sizes (e.g., stage IVA). Indeed, more than 80% of SCLC diagnosed stage was IVB.[6] However, the approach was capable of handling the unbalanced stages avoiding the scenario with all detected subgroups being represented only by IVB stage.

One way of overcoming this limitation is to notably increase the sample size and include longitudinal variables at clinically relevant decision points.

Future studies will focus on the collection of data records of patients diagnosed after 2016, with the inclusion of longitudinal variables. We envision that this will strengthen inference and allow the study of time-dependent effects.

Treatment guidelines are international, and with regard to SCLC, small differences exist between major Western countries. Therefore, the inclusion of data from other centers would constitute an important opportunity to assess the robustness and validity of the approach. Despite the relevant results achieved with the presented approach, different strategies should be explored with changes to the methodologies detailed in the pipeline

(such as covariate selection, survival machine learning, and distance definitions[48–50]). With regard to this, sensitivity analysis will be explored comparing different strategies and methods.

Proper choice of methods can be the key for advancing methodologies aimed to extract relevant clinical information from RWD for risk stratification.[1,2,12,14] This will provide an opportunity to improve our understanding of disease, inform clinical study design, and develop individualized therapies[2,3,13]

ACKNOWLEDGEMENTS

This project is a contribution to the Centre for Data-Driven Health (CDDH), KTH Royal Institute of Technology (https://www.kth.se/cddh).

FUNDING INFORMATION

The Swedish Cancer Society (grant no. CAN 2018/597 and CAN2021/1469 Pj01) to R. Lewensohn and from the Stockholm Cancer Society (grant no. #201202 to R. Lewensohn and #174063 and #201103 to L. De Petris).

CONFLICT OF INTEREST

The authors declared no competing interests for this work.

AUTHOR CONTRIBUTIONS

L.M. and A.S.D. wrote the manuscript. L.M., A.S.D., J.R., and S.M. designed the research. L.M., A.S.D., S.T., A.D., R.L., and L.D.P. performed the research. L.M., A.S.D., S.T., A.D., R.L., and L.D.P. analyzed the data.

REFERENCES

1. Azuaje F. Artificial intelligence for precision oncology: beyond patient stratification. *NPJ Precis Oncol*. 2019;3:1-5. doi:10.1038/s41698-019-0078-1
2. Schurman B. The framework for FDA's real-world evidence program. *Appl Clin Trials*. 2019;28:15-17.
3. Lasiter L, Tymejczyk O, Garrett-Mayer E, et al. Real-world overall survival using oncology electronic health record data: friends of cancer research pilot. *Clin Pharmacol Ther*. 2022;111:444-454. doi:10.1002/cpt.2443
4. Schnog J-JB, Samson MJ, Gans ROB, Duits AJ. An urgent call to raise the bar in oncology. *Br J Cancer*. 2021;125:1477-1485. doi:10.1038/s41416-021-01495-7
5. Tan K, Bryan J, Segal B, et al. Emulating control arms for cancer clinical trials using external cohorts created from electronic health record-derived real-world data. *Clin Pharmacol Therap*. 2022;111:168-178. doi:10.1002/CPT.2351
6. Pietanza MC, Byers LA, Minna JD, Rudin CM. Small cell lung cancer: will recent progress lead to improved outcomes? *Clin Cancer Res*. 2015;21:2244-2255. doi:10.1158/1078-0432.CCR-14-2958
7. Kalemkerian GP, Akerley W, Bogner P, et al. Small cell lung cancer: clinical practice guidelines in oncology. *JNCCN*. 2013;11:78-98. doi:10.6004/jnccn.2013.0011
8. Johal S, Hettle R, Carroll J, Maguire P, Wynne T. Real-world treatment patterns and outcomes in small-cell lung cancer: a systematic literature review. *J Thorac Dis*. 2021;13:3692-3707. doi:10.21037/jtd-20-3034
9. Micke P, Faldum A, Metz T, et al. Staging small cell lung cancer: veterans Administration Lung Study Group versus International Association for the Study of Lung Cancer – What limits limited disease? *Lung Cancer*. 2002;37:271-276. doi:10.1016/S0169-5002(02)00072-7
10. Tendler S, Grozman V, Lewensohn R, Tsakonas G, Viktorsson K, De Petris L. Validation of the 8th TNM classification for small-cell lung cancer in a retrospective material from Sweden. *Lung Cancer*. 2018;120:75-81. doi:10.1016/j.lungcan.2018.03.026
11. Hwang JK, Page BJ, Flynn D, et al. Validation of the eighth edition TNM lung cancer staging system. *J Thorac Oncol*. 2020;15(4):649-654.
12. Booth CM, Karim S, Mackillop WJ. Real-world data: towards achieving the achievable in cancer care. *Nat Rev Clin Oncol*. 2019;16:312-325. doi:10.1038/s41571-019-0167-7
13. Miksad RA, Abernethy AP. Harnessing the Power of Real-World Evidence (RWE): a checklist to ensure regulatory-grade data quality. *Clin Pharmacol Therap*. 2018;103:202-205. doi:10.1002/CPT.946
14. Rivera DR, Henk HJ, Garrett-Mayer E, et al. The friends of cancer research real-world data collaboration pilot 2.0: methodological recommendations from oncology case studies. *Clin Pharmacol Therap*. 2022;111:283-292. doi:10.1002/cpt.2453
15. Breslow NE. Analysis of survival data under the proportional hazards model. *Int Stat Rev*. 1975;43:45. doi:10.2307/1402659
16. Blagoev KB, Wilkerson J, Fojo T. Hazard ratios in cancer clinical trials- a primer. *Nat Rev Clin Oncol*. 2012;9:178-183. doi:10.1038/nrclinonc.2011.217
17. Wang P, Li Y, Reddy CK. Machine learning for survival analysis: a survey. *ACM Comput Surv*. 2019;51:Article 110. doi:10.1145/3214306
18. Gong X, Hu M, Zhao L. Big data toolsets to pharmacometrics: application of machine learning for time-to-event analysis. *Clin Transl Sci*. 2018;11:305-311. doi:10.1111/CTS.12541
19. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat*. 2008;2:841-860.
20. Gower JC. A general coefficient of similarity and some of its properties. *Biometrics*. 1971;27:857-871. doi:10.2307/2528823
21. Kaufman L, Rousseeuw PJ. Partitioning around medoids (program pam). *Finding Groups Data*. 1990;344:68-125.
22. Tendler S, Zhan Y, Pettersson A, et al. Treatment patterns and survival outcomes for small-cell lung cancer patients–a Swedish single center cohort study. *Acta Oncol*. 2020;59:388-394. doi:10.1080/0284186X.2019.1711165
23. Stekhoven DJ, Bühlmann P. Missforest-Non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28:112-118. doi:10.1093/bioinformatics/btr597
24. Genestreti G, Tiseo M, Kenmotsu H, et al. Outcomes of platinum-sensitive small-cell lung cancer patients treated with platinum/etoposide rechallenge: A multi-institutional retrospective analysis. *Clin Lung Cancer*. 2015;16:e223-e228. doi:10.1016/j.cllc.2015.04.006
25. Paumier A, Cuenca X, Le P'echoux C. Prophylactic cranial irradiation in lung cancer. *Cancer Treat Rev*. 2011;37(4):261-265. doi:10.1016/j.ctrv.2010.08.009

26. Takahashi T, Yamanaka T, Seto T, et al. Prophylactic cranial irradiation versus observation in patients with extensive-disease small-cell lung cancer: a multicentre, randomised, open-label, phase 3 trial. *Lancet Oncol*. 2017;18:663-671. doi:10.1016/S1470-2045(17)30230-9
27. Salem A, Mistry H, Hatton M, et al. Association of chemoradiotherapy with outcomes among patients with stage I to II vs stage III small cell lung cancer secondary analysis of a randomized clinical trial. *JAMA Oncol*. 2019;5:e185335. doi:10.1001/jamaoncol.2018.5335
28. Bernhardt D, Aufderstrasse S, Konig L, et al. Impact of inflammatory markers on survival in patients with limited disease small-cell lung cancer undergoing chemoradiotherapy. *Cancer Manag Res*. 2018;10:6563-6569. doi:10.2147/CMAR.S180990
29. Zhang X, Guo M, Fan J, et al. Prognostic significance of serum LDH in small cell lung cancer: A systematic review with meta-analysis. *Cancer Biomark*. 2016;16:415-423. doi:10.3233/CBM-160580
30. Ishwaran H, Kogalur UB. Consistency of random survival forests. *Stat Probab Lett*. 2010;80:1056-1064. doi:10.1016/j.spl.2010.02.020
31. Yan H, Xin S, Ma J, Wang H, Zhang H, Liu J. A three microRNA-based prognostic signature for small cell lung cancer overall survival. *J Cell Biochem*. 2019;120:8723-8730. doi:10.1002/jcb.28159
32. Chen N, Li R, Jiang M, et al. Progression-free survival prediction in small cell lung cancer based on radiomics analysis of contrast-enhanced CT. *Front Med*. 2022;0:292. doi:10.3389/FMED.2022.833283
33. Fisher A, Rudin C, Dominici F. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. *J Mach Learn Res*. 2019;20:1-81.
34. Haas K, Morton S, Gupta S, Mahoui M. Using similarity metrics on real world data and patient treatment pathways to recommend the next treatment. *AMIA Summits Translat Sci Proc*. 2019;2019:398.
35. Yoshimoto T, Motoi N, Yamamoto N, et al. Pulmonary carcinoids and low-grade gastrointestinal neuroendocrine tumors show common MicroRNA expression profiles, different from adenocarcinomas and small cell carcinomas. *Neuroendocrinology*. 2017;106:47-57. doi:10.1159/000461582
36. Liu D, Xu X, Wen J, et al. Integrated genome-wide analysis of gene expression and DNA copy number variations highlights stem cell-related pathways in small cell esophageal carcinoma. *Stem Cells Int*. 2018;2018:1-8. doi:10.1155/2018/3481783
37. Kern JA, Kim J, Foster DG, et al. Role of mTOR as an essential kinase in SCLC. *J Thorac Oncol*. 2020;15:1522-1534. doi:10.1016/j.jtho.2020.05.026
38. Lynch CM, Van Berkel VH, Frieboes HB. Application of unsupervised analysis techniques to lung cancer patient data. *PLoS One*. 2017;12:e0184370. doi:10.1371/journal.pone.0184370
39. Hueman M, Wang H, Liu Z, et al. Expanding TNM for lung cancer through machine learning. *Thoracic Cancer*. 2021;12:1423-1430. doi:10.1111/1759-7714.13926
40. Xu R, Wunsch DC. Clustering algorithms in biomedical research: a review. *IEEE Rev Biomed Eng*. 2010;3:120-154. doi:10.1109/RBME.2010.2083647
41. Ramlov A, Tietze A, Khalil AA, Knap MM. Prophylactic cranial irradiation in patients with small cell lung cancer. A retrospective study of recurrence, survival and morbidity. *Lung Cancer*. 2012;77:561-566. doi:10.1016/J.LUNGCAN.2012.05.101
42. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;24:44-56. doi:10.1038/s41591-018-0300-7
43. Sundstrøm S, Bremnes RM, Kaasa S, et al. Cisplatin and etoposide regimen is superior to cyclophosphamide, epirubicin, and vincristine regimen in small cell lung cancer: results from a randomized phase III trial with 5 years' follow-up. *J Clin Oncol*. 2002;20:4665-4672. doi:10.1200/JCO.2002.12.111
44. Reymen B, Van Loon J, van Baardwijk A, et al. Total gross tumor volume is an independent prognostic factor in patients treated with selective nodal irradiation for stage I to III small cell lung cancer. *Int J Radiat Oncol Biol Phys*. 2013;85:1319-1324. doi:10.1016/j.ijrobp.2012.10.003
45. Hansen O, Sorensen P, Hansen KH. The occurrence of hyponatremia in SCLC and the influence on prognosis A retrospective study of 453 patients treated in a single institution in a 10-year period. *Lung Cancer*. 2010;68:111-114. doi:10.1016/j.lungcan.2009.05.015
46. Rami-Porta R, Bolejack V, Giroux DJ, et al. The IASLC lung cancer staging project: the new database to inform the eighth edition of the TNM classification of lung cancer. *J Thorac Oncol*. 2014;9:1618-1624. doi:10.1097/JTO.0000000000000334
47. Rossi A, Martelli O, Di Maio M. Treatment of patients with small-cell lung cancer: From meta-analyses to clinical practice. *Cancer Rev Treat*. 2013;39:498-506. doi:10.1016/j.ctrv.2012.09.006
48. Oei RW, Fang HSA, Tan WY, Hsu W, Lee ML, Tan NC. Using domain knowledge and data-driven insights for patient similarity analytics. *J Pers Med*. 2022;11:699. doi:10.3390/JPM11080699
49. Sibieude E, Khandelwal A, Hesthaven JS, Girard P, Terranova N. Fast screening of covariates in population models empowered by machine learning. *J Pharmacokinet Pharmacodyn*. 2021;48:597-609. doi:10.1007/S10928-021-09757-W/TABLES/3
50. Terranova N, French J, Dai H, et al. Pharmacometric modeling and machine learning analyses of prognostic and predictive factors in the JAVELIN Gastric 100 phase III trial of avelumab, CPT: Pharmacometrics & Systems. *Pharmacology*. 2022;11:333-347. doi:10.1002/PSP4.12754

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Marzano L, Darwich AS, Tendler S, et al. A novel analytical framework for risk stratification of real-world data using machine learning: A small cell lung cancer study. *Clin Transl Sci*. 2022;15:2437-2447. doi:10.1111/cts.13371